What's fair is. . . fair? Presenting JustEFAB, an ethical framework for operationalizing medical ethics and social justice in the integration of clinical machine learning

JustEFAB

Melissa D Mccradden*
The Hospital for Sick Children
melissa.mccradden@sickkids.ca

Oluwadara Odusi
University of Sheffield Medical School
oodusi2@sheffield.ac.uk

Shalmali Joshi
Columbia University
sj3261@cumc.columbia.edu

Ismail Akrout
The Hospital for Sick Children
ismail.akrout@sickkids.ca

Kagiso Ndlovu
University of Botswana
ndlovuk@ub.ac.bw

Ben Glocker
Imperial College London
b.glocker@imperial.ac.uk

Gabriel Maicas
Australian Institute for Machine Learning
gabriel.maicas@adelaide.edu

Xiaoxuan Liu
University of Birmingham
x.liu.8@bham.ac.uk

Mjaye Mazwi
The Hospital for Sick Children
mjaye.mazwi@sickkids.ca

Tee Garnett
The Hospital for Sick Children
tee.garnett@sickkids.ca

Lauren Oakden-Rayner
Australian Institute for Machine Learning
lauren.oakden-rayner@adelaide.edu

Myrtede Alfred
University of Toronto
myrtede.alfred@utoronto.ca

Irvine Sihlahla
University of Cape Town
irvinesihlahla@yahoo.com

Oswa Shafei
The Hospital for Sick Children
oswa.shafei@mail.utoronto.ca

Anna Goldenberg
The Hospital for Sick Children
anna.goldenberg@sickkids.ca

ABSTRACT

The problem of algorithmic bias represents an ethical threat to the fair treatment of patients when their care involves machine learning (ML) models informing clinical decision-making. The design, development, testing, and integration of ML models therefore require a lifecycle approach to bias identification and mitigation efforts. Presently, most work focuses on the ML tool alone, neglecting the larger sociotechnical context in which these models operate. Moreover, the narrow focus on technical definitions of fairness must be integrated within the larger context of medical ethics in order to facilitate equitable care with ML. Drawing from principles of medical ethics, research ethics, feminist philosophy of science, and justice-based theories, we describe the Justice, Equity, Fairness, and Anti-Bias (JustEFAB) guideline intended to support the design, testing, validation, and clinical evaluation of ML models with respect to algorithmic fairness. This paper describes JustEFAB's development and vetting through multiple advisory groups and the lifecycle approach to addressing fairness in clinical ML tools. We present an ethical decision-making framework to support design and development, adjudication between ethical values as design choices, silent trial evaluation, and prospective clinical evaluation guided by medical ethics and social justice principles. We provide some preliminary considerations for oversight and safety to support ongoing attention to fairness issues. We envision this guideline as useful to many stakeholders, including ML developers, healthcare decision-makers, research ethics committees, regulators, and other parties who have interest in the fair and judicious use of clinical ML tools.

CCS CONCEPTS

• **Social and professional topics**; • **Computing/technology policy**; • **Medical information policy**; • **Medical technologies**;

KEYWORDS

algorithmic bias, fairness, clinical machine learning, ethics, organizational ethics, justice, accountability, healthcare, health policy, safe deployment

ACM Reference Format:
Melissa D Mccradden*, Oluwadara Odusi, Shalmali Joshi, Ismail Akrout, Kagiso Ndlovu, Ben Glocker, Gabriel Maicas, Xiaoxuan Liu, Mjaye Mazwi, Tee Garnett, Lauren Oakden-Rayner, Myrtede Alfred, Irvine Sihlahla, Oswa Shafei, and Anna Goldenberg. 2023. What's fair is. . . fair? Presenting JustEFAB, an ethical framework for operationalizing medical ethics and social

FAccT '23, June 12–15, 2023, Chicago, IL, USA

ACM ISBN 979-8-4007-0192-4/23/06…$15.00
https://doi.org/10.1145/3593013.3594096

justice in the integration of clinical machine learning: JustEFAB. In *2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23), June 12–15, 2023, Chicago, IL, USA.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3593013.3594096

1 INTRODUCTION

"If inequity is woven into the very fabric of society then each twist, coil, and code is a chance for us to weave new patterns, practices, politic. Its vastness will be its undoing, once we accept that we are pattern makers." Dr. Ruha Benjamin

Artificial intelligence (AI)/machine learning (ML) systems are recognized as potential sources that worsen societal inequities through algorithmic bias. 'Algorithmic bias' refers to the unequal performance and disparate impact of computational systems utilizing AI methodologies. In the clinical sense, algorithmic bias can appear as disparities in the performance (e.g., accuracy, error rates, true/false positives and negatives) indexed to gender, sex, race, ethnicity, language, socioeconomic status, and other identities which are not indexed to clinical need. By virtue of this relatively worse performance among specific groups, ML can exacerbate health disparities by reifying an unfair status quo [1]. This disparate performance implicates traditional ethical principles of justice, which is concerned with the treatment of individuals and/or groups. Abeba Birhane refers to these impacts as 'algorithmic injustices' (drawing broadly from social justice principles) whereby the use of the algorithm reinforces, reifies, and/or exacerbates existing inequities [1]Relatedly, the field of 'Fair ML' has been developed in response to the identification of algorithmic bias as a phenomenon of ML approaches [2] and has sought to identify practices whereby a ML model can be defined as 'fair' [3] [4]. These efforts are important to both the detection and mitigation of algorithmic bias as a technical problem for healthcare ML.

Algorithmic bias can result in both direct and indirect disadvantages. Direct disadvantages (i.e., through computational performance itself) occur due to lower accuracy, greater error/failure rates, or more uncertainty of predictions. Indirect disadvantages (i.e., secondary to algorithmic performance) occur when the prediction may be actuarially correct, yet result in problems such as reducing the quality of decision-making by clinicians receiving algorithmic predictions, disparate allocation of resources as a function of one's identity rather than indexed to clinical need, or operational impacts such as the redirection of scare healthcare resources. Both call into question the need to understand the distributive justice properties of a given system – the relative distribution of benefits and risks – in order to promote fair treatment of persons when using ML systems.

There is a need to connect the intentions of the Fair ML work with canonical conceptualizations of justice in the medical ethics literature in order to promote ethical ML in healthcare. This guideline takes a position recommended by many knowledgeable scholars [1] [5] to avoid the "veneer of technical neutrality." It is acknowledged that ML systems rely on imperfect data and are predisposed to approaching problems in idiosyncratic and often inhuman ways, relying on factors in their decision-making processes which are opaque to human users. Nonetheless, we must make decisions; it is apparent from an unfortunate plethora of examples that without due consideration of bias, systematic discrimination and other injustices can occur when models built from such imperfect data are used for decision-making. Moreover, when so-called 'ethical' solutions to bias are employed without sufficient knowledge and attention to health equity literature and community engagement, disparities remain and can even worsen [6] [5].

This guideline does not purport to offer 'solutions' to fairness. By identifying the ethical strengths and vulnerabilities in support of a particular scientific endeavour, we can make our logic for ML development transparent, open to scrutiny by scholars and publics, and justify our choices around integration ethics rather than statistics alone.

1.1 Relevant prior work

1.1.1 Characterization of bias in Health ML.. There have been several efforts to systematically detect and characterize algorithmic bias in healthcare machine learning [7] [6] [8] [9] [10]. These have included practices ranging from algorithmic audits [11] [12] [13], documentation practices [14] [15], impact assessments to direct algorithmic designs and amendments intended to address and implement fairness [16] [17] [18]. Some works provide over-arching principles for addressing fairness in healthcare systems [19] [20] [21] [22] [23] [24] [25].

1.1.2 Computational strategies for algorithmic bias mitigation. There are three commonly identified categories of algorithmic bias mitigation strategies which may be implemented to improve model fairness by modifying the training data, learning algorithm, or predictions. Several resources exist describing preprocessing, in-processing, and post-processing methodologies (as a non-comprehensive list, see: [4] [20] [3] [26] [27] [28] [29] [30] [31]). There are no current consensus-based standards for applying these methods.

1.1.3 Operationalizing fairness for ML in healthcare. A gap in the literature pertains to the connection between ethics-related ML design choices (like algorithmic fairness methods) and on-the-ground evaluative practices that considers the clinical context and moral plurality of the various environments in which ML might be utilized. Primarily, algorithmic operationalization of fairness has focused on performance differences between models in conjunction with post-hoc re-calibration strategies. On the other hand, a prediction itself constitutes a smaller portion of clinical decision-making, which involves accounting for clinical context, patient preferences, in conjunction with other ethical considerations [32]. As a result, most works in fairness in ML for health have argued for operationalizing a much broader view of bias mitigation as opposed to narrow performance-based fairness considerations [3] [33]. That said, performance-based fairness evaluations [24] [23] [34] [20] [19] are crucial as a component in the design process of an ML model and provide an important computational basis upon which to build out to consider the larger ethical context.

Initial guidelines have primarily focused on reporting and publishing ML-based models in health, without substantive discussion of the fairness-related issues pertinent to these models [17]. For example, considerations of algorithmic fairness approaches typically approach 'fairness' by focusing on the ML model in isolation, separate from the clinical task or environment. As such, these works

often focus on making *predictions* 'fair' by one of the many technical definitions, with the determination of what is fair being defined (typically) at the discretion of the ML researchers [21] But a 'fair' algorithm in aggregate does not ensure fairness to the individual who is the recipient of the technology; only that the algorithm has satisfied a given technical definition [33]. Similar to how technical accuracy is not a guarantee of patient benefit [35] technical fairness is not a guarantee of the fair treatment of individuals.

1.2 Aim

The aim of this paper is to describe SickKids ethical decision-making framework to address fairness issues in healthcare ML tools affecting patient care across the lifecycle of ML design, development, validation, integration, and oversight. The guideline extends prior work in algorithmic fairness by connecting ML performance to the intended impact on patient outcomes as well as identifying adjunct practices to support an overall sociotechnical approach to fairness.

2 METHODS

This framework was developed as an institutional policy at The Hospital for Sick Children (SickKids) that we tested for generalizability. Mapping onto a lifecycle approach to AI integration [36] [37] stages include 1) design and development; 2) in situ evaluation (silent trial evaluation); 3) prospective clinical evaluation; 4) ongoing monitoring. Content was developed initially by the lead author in conjunction with collaborators and iterated upon across multiple use cases; it continues to be a 'living document' open to further refinement with an eye toward robustness and adaptability to different applications of ML. It should be noted that as an institutional policy, we do not strive for reproducibility as the development is highly local in nature; rather, we aim to comprehensively describe our process toward the aim of transparency in the steps undertaken for this guideline. The complete framework is presented in Appendix A, the use case application in Appendix B, and complete Methods in Appendix C.

3 THE JUSTEFAB FRAMEWORK

JustEFAB takes the position that ML models are sociotechnical tools [38]. To act ethically, we must combine good technical choices with the established norms that characterize the people who are the intended recipients and users of the technology. This positioning means that to ethically design, test, and implement a given tool, we must work backward from the implementation environment's norms. The established norms of medicine include the principles of biomedical ethics and clinical research ethics. Increasingly – and most relevant to the issue of algorithmic bias – medical pedagogy is incorporating social justice principles in recognition of underserved and mistreated groups. As such, we consider the application and relevance of these principles across the lifecycle of ML integration. The framework's ethical guidance is described throughout, including what these principles and theories call on us to do for ML in health. The framework is presented in Table 1. (Appendix A).

3.1 Guiding theoretical commitment

The overall approach to this guideline was guided by feminist philosophy of science, and standpoint theory in particular. Standpoint theory refers to a general orientation toward the core tenets of standpoint epistemology (recognizing that there is a diversity of views within scholars who are standpoint theorists). These core tenets include the notion that knowledge is socially situated [39] [40]; people are marginalized can be epistemically advantaged in others ways with respect to knowledge that is lacking among dominant perspectives [41] [40]; standpoints are the result of work, and are not simply granted by virtue of one's identity [42] [41] [40]; centering marginalized perspectives provides a more accurate means of illuminating social phenomena [41]. For AI, standpoint theory suggests that we ought to frame fairness analyses from the informed perspective of marginalized groups [1], seek out knowledge regarding the implications of ML unfairness where relevant, and integrate multiple perspectives to better understand the structural issues for ML integration.

An additional component of standpoint theory is the appreciation of multiple ways of knowing. For example, ML outputs may be considered one form of knowledge that contributes to a larger picture [32]. Ethically, it is important to consider the ways in which multiple forms of knowledge come together to form the decision-making picture [43]. Moreover, ML tools which may be used across cultural and legislative contexts will need to reconcile values in tension to realize a shared goal of using new technologies to enhance car [44] Shared decision-making recognizes that patient values, evidence-based medicine, and clinical judgment together form the basis for good decisions. We note that from a justice perspective, further steps are needed to integrate, for example, Indigenous ways of knowing to better respect these individuals and communities and reduce health inequities [45] [46]. This was a key point stressed by our consultative groups.

3.2 Stage 1A: Design and Development

Given historical under-attention to concerns of fairness more generally in both healthcare and medicine, we advocate for testing every model that will directly influence patient care decisions. This commitment means that at SickKids, every single model should undertake a fairness analysis. Drawing from sources such as [24] [23] [34] [20] [19] we outline the following steps.

3.2.1 Preparation and conceptualization. A critical recommendation at this stage is to conduct a literature search or formal review to identify the current stage of knowledge regarding potential health inequities implicated in the model's tas [23] It should be noted, however, is that the absence of documentation of disparities does not mean that no disparities actually exist. Characterizing these gaps is a necessary step toward equitable care delivery and stakeholders should not be deterred from documenting health disparities. Nonetheless, describing the current state of knowledge is an important starting point and there are the limitations of drawing from the traditional evidence hierarchy. Randomized controlled trials (RCTs) are among the strongest forms of evidence, but have suffered from limited inclusivity with respect to marginalized groups [47] [48] [49] [50]. Social science and qualitative research can contribute additional insights into problem formulation and understanding of data and patterns therein.

The literature review offers several opportunities to improve the scientific and ethical rigour of model design. By identifying health

inequities, the developers can identify a priori which groups for whom fairness properties should be evaluated. The identification of such groups can also provide direction for consultation regarding implementation; for example, by consulting scholars knowledgeable on the intersection of race and medicine, the developers can better establish the problem formulation [23] Such consultation may prevent future algorithmic discrimination [5]

A priori group identification: When considering algorithmic bias, a key element is that its performance discrepancy is systematic, meaning it applies broadly to a given group. While algorithmic performance can vary significantly according to many different patient features (e.g., presence or absence of a chest tube; [51]not all of these differences are considered 'unfair.' To determine whether a performance gap is unfair, we must first consider how patient groups are defined. A 'group' can be considered as "a collective of persons differentiated from at least one other group by cultural forms, practices, or way of life. Members of a group have a specific affinity with one another because of their similar experience or way of life, which prompts them to associate with one another more than with those not identified with the group" [52]. By defining groups in this way, we can identify on what basis the differential impact of algorithmic performance should be assessed.

Problem formulation: Ziad Obermeyer's influential work has documented how varying the problem formulation can result in different implications for fairness [6]. An algorithm trained on total healthcare expenditure can demonstrate substantial algorithmic discrimination against Black patients when the same data trained on total medical visits, clinical problems or others can result in equal impacts to Black and White patients [6] The problem formulation is the ideal time to engage resources like ethics, equity diversity and inclusion groups, community partners, patient stakeholders, and others as many of these groups can quickly identify ethical challenges to problem formulation.

Consultation: The identification of the groups deemed most at risk from algorithmic bias is a means of identifying those with whom consultation is warranted. Standpoint theory posits that membership of a group alone is insufficient – individuals must have developed knowledge and understanding of the power dynamics and structures that relate to the apparent disadvantages among that group [42] [41] [40]. Knowledgeable individuals and groups can shed light on important context around the labels themselves, the understanding of patterns of inequities, and indicate a desirable state that could be pursued by the use of the algorithm. Note that problem formulation can and should be re-visited after a characterization of the performance of the candidate model.

3.2.2 Dataset inclusivity. The importance of dataser characterization as it applies to AI and healthcare is that without sufficient high quality data training models will be biased and not generalize appropriately. Yet, in medicine, data as a substrate for learning about and understanding causally-relevant patterns meant to generate insights into patients' conditions, diseases, and prognoses is highly vulnerable to structural problems that compromise its quality [53]. Assessments of data being 'fit for purpose' is a valuable component of good ML practice and can inform decisions made regarding bias patterns [53]. In some cases, the fitness of the data will be unsuitable to the proposed model integration plan from a fairness perspective, and so the research team may decide to either change course or abandon the model. Knowing that we are working with data that is an imperfect representation of the medical phenomenon one wishes to model enables reflexivity and epistemic humility.

Representation and labels: Based on the original Datasheets for Datasets [14], Healthsheets for Datasets [15] provides a resource for developers to reflect on their data. Anecdotally, we find that for many healthcare institutions, more work is needed to better describe the dataset properties [54]) and we anticipate further work in this area. Systematic differences in data representation can occur at the level of both data sources and population [20] Data sources (e.g., electronic and administrative health records, clinical trial and research data, and social media) can contain systematic variances in patterns of representation depending on structural issues like access. Population-level issues can arise when a given group is less well represented either numerically or in data quality. Identification of under-representation (as a matter of individual and multi-group identifiers) at this stage can contribute to accuracy and equality by highlighting the groups where more information is needed to provide better algorithmic outputs. For example, in dermatological datasets, representation from more diverse skin tones has been proven to improve algorithmic performance at the identification of various skin cancers [55]. In some cases, label selection can be challenging when lacking attributes like race and/or ethnicity, as is the case for many places such as Canada. As novel methods are developed and utilized in healthcare (e.g., federated learning [56]), the ability to analyze for fairness remains critically important [57].

Chen and colleagues similarly note that outcome definition can be a source of bia [20] Systematic differences in assigning clinical diagnoses, labels, or assessments of risk can fall across a spectrum of objectivity. For example, ML-based detection of cardiac rhythm abnormalities may be relatively more objective than assigning diagnoses of schizophrenia or alcohol use disorder. Similarly, documentation in patient charts may reveal such biases; it is well recognized that marginalized patients are labelled in certain ways which are documented on the chart (e.g., 'aggressive,' noncompliant,' 'drug-seeking'). Therefore, exploration of these sorts of problematic labels as data inputs can also be valuable to the process of elucidating algorithmic bias. From these assessments, decisions can be made about how to manage these problematic labels as consistent with the ethics-oriented goal for algorithm design. To explore options, algorithms can be retained on different labels and compared directly to assess implications for bias resulting from the available model option [23] An additional option may be to collect new data directly to improve the model [23] Finally, some data may be so problematic that the proposed model task should simply not be pursued.

Data reflexivity: A priori group identification provides a means to consider how the labels in a given dataset may or may not represent the groups of patients between whom one must test for model performance. The concept of intersectionality [58] highlights how marginalization can intersect across different identities to produce relative levels of advantage and disadvantage. For example, race, sexual orientation, gender identity, immigration status, socioeconomic status, dominant language fluency, and disability are all dimensions upon which privilege and marginalization may intersect to influence a person's social status. An acknowledged limitation of algorithmic fairness approaches (and indeed this guideline as well)

is that current methods for assessing bias take only one feature in isolation and compare to other groups (e.g., Black versus White patients). Comparing algorithmic performance according to individual identifiers is thus fundamentally at odds with the formal theory of intersectionality [58]. Some authors have sought to follow the spirit of intersectionality and apply it to ML [59] [60]. At the same time, we the need to statistically control for covariance, which is more difficult when computing multiple identifiers and is best done by identifying distinct groups.

Analytic plan: Keeping these caveats in mind, it is nonetheless vital to use some form of classification of individuals to better understand the implications for a given ML application's performance. The most important issue to stress for this guideline is that selection of groups for fairness (and, later, outcome) evaluation is necessary to advancing equity. Labels are inevitably proxies, typically over-simplified; so, deciding and documenting the rationale for label selection aids with transparency of the model's evaluation and can guide clinician users in their interpretation of its outputs. For example, knowing that a model's performance was analyzed by sex as an identifier on a health card helps clinicians to readily identify its limitations with respect to trans and gender diverse patients.

A caveat must be noted here. Inevitably, the metrics we use to compare groups are generally proxies for the factors we consider to be truly relevant for the prediction problem. Many advocate for race-based data collection, for example, to correctly label individuals for group-level comparison. While race-based data is valuable in several respects, it is important to also keep in mind that these labels themselves are social constricts which can be imprecise and ontologically confused [61] [62] [63]. The United States, for example, typically categorizes individuals as 'Black,' 'Asian,' 'White,' or 'Other.' These labels offer limited insights into the person's racial identity, ethnicity, or heritage; neither do they offer a quantification of adverse events such as racism which are direct mediators of the health outcomes being studied [64]. In other cases, eliding race and ethnicity also compromises the scientific quality of the prediction task - for example, treating 'African American' as a singular ethnicity neglects the genetic diversity across persons of African descent [65]. In many cases, 'race' is considered to be a proxy of racialization from a fairness perspective. But this logic cannot be applied across the globe, and will be subject the patterns of fairness exhibited locally. Scholars have noted the general dominance of Western/Global North perspectives on fairness which further underscore the need to reflect on and consider data labels in their unique contexts, based on the health needs locally, reflecting the diversity of the population the model will serve [66].

Similarly, dividing patients into 'male' and 'female' belies the extensive heterogeneity that exists within each group, in terms of physiology, hormones, and experience (e.g., gender-based violence) [67] [68]. A precision medicine approach would seek to quantify features that directly and indirectly modify the disease course (e.g., hormone levels) rather than relying on imprecise labels such as male/female. In the EHR, misgendering, pathologization, and medicalization of 2SLGBTQIA+ and gender diverse persons is reflected in the labels, features (e.g., in clinical note modelling), and patterns of care that can be reflected in ML tools. Data quality for trans and gender diverse persons can be low given their distrust and hesitation in medical settings stemming from a high frequency of negative experiences in healthcare settings [69] [70] [71]. For some excellent resources on gender annotation in the EHR, the reader is referred to [72] [73].

3.2.3 Algorithmic validation. ***Test and compare performance between patient subgroups***: The next step is to analyze and explore bias by comparing groups in terms of the algorithm's dis-aggregated performance, beginning with those identified a priori based on the initial review and consultation [23]. Again, we stress that this analysis should be done locally in response to the groupings relevant in that context. Identifying whether or not a pattern of disparate performance is apparent starts with comparing metrics like accuracy, error rates, and uncertainty [23] to begin to characterize the fairness properties of the candidate model(s). Statistically significant bias must be assessed while controlling for confounding. Base rate differences in performance may be influenced by the underlying characteristics of the groups being compared. For example, Seyyed-Kalantari and colleagues [10] document performance discrepancies of a diagnostic system across multiple patient identifiers. Commentators note, however, that that many differences observed can be wholly explained by systematic differences (e.g., age, disease severity) between groups [74] [75]. Sampling error can occur when individuals represented within datasets are not representative of the whole patient population that the model is intended to address. These samples can be missing at random or missing systematically; there are different implications for random versus non-random missingness. In addition, systematic measurement error can occur when the labels associated with particular identities are more error-prone. We consider 'ethically significant bias' as a situation wherein the disparate performance of a model could have systematic negative consequences for the treatment of a particular group of patients based on their identity, and that this treatment is not reflective of clinical circumstances. The silent trial offers a means to more reliably test for and hypothesize about the potential consequences to patients.

Importantly, it should be noted that in many cases the model's input data will reflect extant influences irrelevant to the prediction task – meaning, that data is nearly always 'biased' in the sense that it reflects some amount of unfairness. Because of these patterns, a fair model can be perceived as unfair when evaluated on biased data; similarly, an unfair model can go undetected when evaluated on unfair data. Part of the problem lies in the assumption that the same biases will be apparent in the training and test data. This limitation underscores the importance of the silent trial as the more reliable source of evidence regarding fairness patterns.

Algorithmic fairness methods: A complete review of algorithmic fairness methodologies is beyond the scope of this paper. The reader is referred to a number of excellent resources, including [18] [34] [76] [3] [26] [4]. For the purpose of this framework, we stress that choices about fairness methodologies employed should be: 1) *informed* by strong knowledge about the nature of the health inequity influencing the main prediction task; 2) *evaluated* as part of an algorithmic validation process in addition to evaluating the on-the-ground performance; 3) *revisable* as supported by an ongoing monitoring process [77], and subject to modification based on the clinical evidence observed through real-time model use

We also stress that the most upstream correction should be taken wherever possible. For example, improving the data quality is preferable to adjusting for poor quality data. We consider the use of algorithmic fairness methods to be legitimate wherein their use would facilitate a beneficial change to the treatment of patients were the outputs to be directly actioned upon. For example, an algorithm that is adjusted to facilitate more referrals to specialist care for a historically under-served group would be considered legitimate to the point where the actual referral rates become reflective of the clinical need across the relevant populations [78]. Decisions about the legitimacy of algorithmic fairness methodologies should be based on the actual clinical impact they have to affected populations, which will be apparent through prospective testing.

Post hoc testing (hidden stratification): In addition to a priori group-based comparison, post hoc methodologies can further inform understanding of the fairness properties of models. It has been observed that there can be clinically meaningful subgroups for whom model performance may differ, known as 'hidden stratification' [51] Methodologies exploring the model's properties as a whole (inherent explainability) can be useful.

In some cases, it may be important to test whether there is a signal in the data that is predictive of subgroup association without using the specific identifier. Banerjee and colleagues [79] identified how deep learning models are capable of identifying patient race from image data alone. By identifying whether there is a signal in the data pertaining to a particular group, developers can identify the degree to which a singular identifier may or may not contribute to performance discrepancies.

3.3 Stage 1B: Ethical decision-making

Based on the information gathered in 3.1.3, informed decisions about model design can be made – ideally, these are made in collaboration with consultants and taking a multidisciplinary approach.

3.3.1 Identifying ethically significant biases. The decision about whether an algorithmic bias is significant is ethically significant implicates our values and beliefs about what constitutes 'fairness.' For example, prediction of sex-linked conditions (e.g., colour blindness) will be 'biased,' but this bias would not be ethically significant because the natural prevalence of the condition is not unfair in itself. This judgment requires strong knowledge of the condition, as indicated in 3.1.1. The bias in performance discrepancy should be analyzed through the lens of biomedical ethics and social justice (NB: this lens can be adapted locally based on established norms). As above, 'ethically significant bias' we define as a discrepancy in an algorithm's performance which would result in systematic differential treatment of patients based on identity, not indexed to clinical need.

3.3.2 Reflective equilibrium. Reflective equilibrium is the process by which moral agents reflect and weigh the facts against the principles relevant to a given case with the goal of providing a robust justification of the selected action [80] [81]. The goal is to identify a proposed strategy, its justifiability, and its risks in relation to the larger goal. Reflecting across principles of medical ethics, local values, policies, professional guidelines, and relevant law can help to inform the analysis. The reasoning behind the selection is important as it provides a means to justify and trace back the choice that was made – its ethical strengths and vulnerabilities.

3.3.3 3.2.3 Ethical choices. Having identified the group(s) for whom algorithmic bias is present, we consider what the ideal pattern would be in terms of patient outcomes to identify what strategies are possible for achieving this ideal state. This thinking can and should be informed by the literature, by knowledgeable stakeholders and scholars with specific experience, and by clinicians who know the context in which the tool will be used. By thinking about the ideal stage, we can consider how the use of the algorithm and what methods specifically can help us to change practice. Finally, we consider which outcomes should be evaluated to determine whether and to what extent we have achieved that aim. We observe generally three categories of guiding values in this regard: predictive accuracy, formal justice, and distributive justice.

Predictive accuracy: 'Accuracy' refers to a model for which actuarial accuracy is the priority, including, potentially, the influence of unfair patterns on its predictions. Prioritizing accuracy is generally the status quo for most model implementation, given the value of computational systems providing actuarially correct information to decision-makers [82]. These accurate predictions, however, may still be influenced by bias relating to structural and social inequities.

This option is ethically supported when there is no tangible benefit to patients that would be gained by modifying the algorithm's functioning or output. Typically, this happens when the causes of unfairness are outside of the control of the clinician (e.g., by structural factors). The prioritization of accuracy over other ethical values is most reasonable when a) accurate predictions are primary to clinical value and b) the care of a disadvantaged group will not be improved (in terms of outcomes) by adjusting the algorithm or outputs. For example, childhood asthma prevalence is related to socioeconomic status. The likelihood of a child having asthma is related to their exposure to environmental toxins among other factors. Correcting for this bias by down-weighing the influence of this sensitive attribute (socioeconomic status) may introduce harm by lowering the potential capture rate of asthma among such children. To the contrary, this option would not be justified where prevalence is influenced by prejudice; for example, Chasnoff [83] reported that birthing women who were Black were 10 times more likely to be reported for drug use, despite similar rates of alcohol and drug use between Black and White pregnant women. To model accuracy in this case means modelling a prejudicial and biased practice of detecting problematic drug use, which would reify harms to these individuals.

Accuracy can feel ethically incomplete when health disparities remain. By virtue of having revealed a discrepancy between the predicted pattern and an ethical ideal, we may consider how to fill the gap. Developers or healthcare decision-makers may wish to include information either during the training of clinical users or accompanying the model outputs to support ethical use of the model. For example, drawing attention to unrecognized needs, inclusive language, and cultural humility alongside model predictions can contribute to the overall improvement of the healthcare environment independent of the model. Additionally, although some corrections are beyond the remit of developers, the latter can play

a role in signalling to clinicians the important limitations of the model and bring forward any suggestions from their consultative work as part of the model's development. Non-ML strategies for bias mitigation are highlighted in 3.3.2. As an example, consider the dermatology case wherein increasing representation of cases of dark-skin individuals improves model performance [55]. While an important step to equitable care in dermatology, disparities in outcomes may still be driven by under-appreciation of the problem of skin cancer among the same individuals, which can lead to delays in accessing and receiving care and thus influencing outcomes [84].

Formal justice: Formal justice originated with Aristotle and follows the common intuition of 'treating like cases alike.' There are distinctions in the philosophical literature as to whether equality may be conferred via a fair process versus fair outcomes [81]. For a given ML application, we can consider whether the predictions themselves by virtue of computational process (e.g., algorithmic fairness methods) can promote the equal treatment of individuals (fair outcomes) or whether the calculations themselves are substantively fair, meaning that they take into account causally relevant patterns and minimizes or eliminates extant influences on predictions.

The rationale for formal justice as justification also depends on the nature of the prediction task. To continue with the above example of asthma detection, the same rationale might not apply to the allocation of resources for asthma care. Knowing that marginalized children are less likely to get access to specialist care, if the model's task were to allocate resources then faithfully replicating the on-the-ground pattern would reinforce disparities in access. A model aiming to advance fairness of outcomes would be designed such that predictions aim to operationalize clinical need rather than past patterns of access which are influenced by societal unfairness (e.g., recall [6]).

A more complicated situation is that of prognostication predictions. As severity influences disease trajectory and prognostication, it is possible that given the structural influences on racialized children, they are more at risk of poor outcomes – a factor which cannot be ignored by the clinician endearing to provide accurate information to families to make decisions. However, it is important to also consider the potential influence of structural biases on the knowledge we have pertaining to different medical issues. For example, the perception of disability as a universally 'bad' outcome has negatively (and inaccurately) influenced prognostication of severe neurological injuries [85].

As with accuracy, there are times where equality will be ideally supplemented by additional measures to support ethical use of the model. As above, decision-makers may wish to consider what information should accompany the training and outputs to best support ethical model use. Again, while beyond the remit of ML developers, signalling residual biases in the model's prediction patterns can be a valuable piece of information to support ethical clinical decision-making and aid clinicians in recognizing the consequences for vulnerable patients [22] [33].

Distributive justice: Another way of promoting fairness is by advancing the needs of the least well-off, known as distributive justice [81]. Rawls argued that treating individuals differently can be justified over equality if such treatment effectively improves the well-being of those with relatively fewer material advantages. Distributive justice should be considered when equality will simply enforce a status quo that is considered unfair and contributes to differential treatment. For example, Park and colleagues identify that Black birthing parents have under-recognized mental health needs postpartum and compare algorithmic fairness methods for detecting postpartum depression (PPD) in the context of under-referral of this group for care [78]. They remark that some methods can improve fairness in detection of PPD (and thus facilitate care referrals) without compromising the model's accuracy, while others involve a trade-off between fairness and accuracy. Although it is indeed likely that the actuarial accuracy of the model may be compromised by enforcing a definition of fairness, it is possible (and maybe even very likely) that the prospective evaluation of a less accurate model adjusted to improve sensitivity of PPD among Black parents could show strong *clinical* accuracy. In other words, if the on-the-ground situation is that Black parents have under-detected PPD, and this pattern is apparent in the unadjusted algorithm, then adjusting the algorithm to specifically improve detection in this group could be more true to reality. This case again highlights the need to thoroughly consider the evaluative scheme for the model's prospective performance with respect to not just predictions but clinical outcomes. Additionally, the feedback loop to the model provides another opportunity to reflect on these design choices and consider the clinical evidence with respect to improving health disparities.

As with all options, some residual ethical considerations remain. If individuals do not feel psychologically safe or are effectively unable to access care, then outcomes will not reflect an improvement to a health disparity despite the model being 'fair.' To truly improve equitable care, we need to ensure that culturally and psychologically safe, accessible systems are in place for patients and families.

3.4 Prospective non-interventional evaluation: the silent trial

3.4.1 Clinical performance. ***Clinical accuracy***: Once the model is validated and deemed a candidate for translation, a prospective non-interventional (silent) trial may be conducted to establish the ecological validity [36] [86] [87] [88]. This step enables an on-the-ground assessment of the model's clinical performance and feasibility without yet impacting patient care. Though anecdotally considered a 'sanity check' for a model's on-the-ground performance, the role of the silent trial may be much more important from an ethics perspective. Unreflective implementation of models without a fulsome preclinical evaluation is tantamount to research waste, violating the requirement of social and scientific value [89] and risking future trust and acceptance of beneficial ML tools. Silent trial evaluation also provides a mean to identify material harms that may arise from algorithmic biases and inappropriate tool use [88] prior to the point when it is affecting patient care. Even where participants in prospective clinical trials agree to take on risks for the purpose of advancing scientific knowledge, it is ethically desirable to embed upstream processes that can minimize these risks.

Characterizing performance across subgroups: At this stage, one can determine whether the design choices made on an ethical basis are reasonable from an implementation perspective and hold the potential to improve care by examining the on-the-ground

model performance with respect to the subgroups identified in 3.1.1. Note that here again the analytic plan must take into account the statistical considerations relevant to support the identification of performance discrepancies across different groups. The sample size calculation will need to take these into account.

Auditing: Following Raji et al [12], some scholars have sought to establish best practices around algorithmic auditing for healthcare ML models [11] [13] [90] [86]. These practices are consistent with clinical trial reporting guideline recommendations to characterize failure cases and failure modes of health AI systems [91] [92]. The practice of auditing provides a robust and comprehensive means of characterizing a model's performance overall. This information can be relevant to research ethics review, regulatory bodies, and clinician users to better understand, as a whole, the performance of the model. Algorithmic errors and failure modes may come to be embedded into patient safety mechanisms and adverse event reporting.

Revision: From the results of the silent trial, stakeholders can review their choices made in 3.2.1 regarding the intended goal for improving a health disparity. If there are unexpected results, it may warrant returning to problem formulation, consultation, or changes to the model's design.

3.4.2 Human Factors. ***Human-centred design***: An increased focus on human factors has been stressed in the literature [93] [94], with human-centred design a commonly acknowledged value in support of responsible integration. This work can be conducted at the silent trial stage where the implementation/integration aspects at the point of the silent trial become more salient and there is still the opportunity to revise the model's integration strategy prior to actual clinical use. Though a full review of human factors in ML is beyond the scope of this work, we highlight the implications of human factors for JustEFAB.

It is widely recognized that end users (typically, clinicians) should be engaged to support design choices. We have stressed consultation through this framework, and again turn back to consultation as an important practice to support ethical integration. Avoiding tokenism and 'rubber-stamping' by these groups is important to meaningful engagement, including with patient and family partners [95].

Integrating equity, diversity, and inclusion (EDI) values into user research (e.g., inclusive sampling practices), user interface requirements, and use-related risk analysis are important to preventing engineering toward the 'dominant group.' Inclusive sampling practices are discussed further below. User interface requirements must consider issues like colour blindness, sensitivity to light levels, and other ergonomic factors can promote usability of the tool. From an ethics perspective, these design elements support inclusivity of the model; if persons with disabilities have barriers to using the tool, they are likely to feel excluded. Use-related risk analysis is supported by the considerations outlined above with respect to defining subgroup-specific performance, and may also be considered with respect to user interpretations of the interface.

Respect for persons: Much skepticism about ML model use in healthcare stems from the historical dislike of interventions which were implemented for use without the desired level of input from stakeholders. For example, electronic health records are widely recognized as a value-add for patient care, yet are consistently maligned by HCPs [96]. The act of meaningfully engaging the intended user group in the design and development is a demonstration of respect for their lived experience as individuals, the knowledge they hold from having this experience, and of the values they hold in doing their work. However, tokenistic engagement can entirely reverse the intended benefits.

Ethical use considerations: Considering the information needed to support clinicians' understanding of the model's development and clinical reliability can also promote ethical use as well as form an augmentation to ML integration [33]. For example, the use of language on the interface holds power. It can influence how users use the predictions and the tool. Considerations for avoiding stigmatizing language, vetting language for racially- or gender-coded sensitivities, and other review with an eye to inclusion will best support ethical use. An additional opportunity to embed ethical considerations is in the training of clinicians – for example, ensuring clinicians are equipped to use the tool appropriately to over over-trust and center patients' interests. Clinicians should be provided with the specific features that were used for a fairness analysis so that they may exercise their judgement in applying the algorithm's output to an individual patient.

3.5 Prospective clinical evaluation

The prospective evaluation of model performance and its impact on care delivery is paramount to judicious use of ML. The model's clinical performance can be influential to determining how much or little its predictions influence clinical decision-making [32]. Here we include important considerations for JustEFAB in the prospective evaluation of models.

3.5.1 The importance of diversity in AI clinical trials. Diversity in clinical research representation is a matter of great interest more broadly than for ML alone. As noted above, it is important to be specific about the value of diverse representation in clinical trials involving AI to avoid notions of biological essential of race and gender. Without adequate understanding of the complexities of identities such as race and gender and their relationship with outcomes, we risk recapitulating, reifying, and exacerbating dangerous stereotypes [97].

Clinical trial reporting has previously stressed the importance of reporting outcomes among different patient subgroups. Similar considerations as we note above concerning label choice and patient identifiers should be noted caveats to outcome reporting as well. Statistically, there is an issue with multiple comparisons - as the number of groups being compared grows larger, the reliability of statistical tests of difference grows smaller (particularly for those concerned with intersectionality). We therefore stress that it will be necessary to have some amount of grouping, such that we can expect that some meaningful differences between individuals are camouflaged by subgroup labelling. These concerns can be mitigated through the education and training of ML users.

We consider the notion of diversity in clinical trials as one of an equality of opportunity: every person should have the opportunity to participate in research for the purpose of advancing scientific knowledge. Trial design should be inclusive to enable diverse representation so that persons experiencing structural vulnerabilities are

not prohibited from participation. A careful eye to balancing compensation and benefit from research participation against coercion can be struck to preserve the voluntariness that is quintessential to ethical research [98]. To improve inclusivity in clinical research requires a much more multidisciplinary effort. Consultation (as above) can elucidate trustworthy practices to support inclusive research.

3.5.2 Consent bias and under-represented (marginalized) individuals. It is commonly believed that certain groups (e.g., Black, Indigenous, Native American, and Aboriginal persons, in particular) are less likely to consent to participate in research. There is certainly an over-representation of White participants in many areas of research, with relatively lower participation among racialized and other oppressed groups. Some even argue that there is a 'consent bias' and so to promote 'fairness' we should stop asking for consent.

There are two problems with such statements. First, recent research shows that the problem of under-representation may be at least partially driven by a failure to approach non-White patients and families to assess interest in participation in research. For example, a recent study of eligible versus enrolled biobank patients demonstrated that determining eligibility was a major driver in representation, rather than being a matter of consent alone [99]. Anecdotally, it is well known that clinicians exercise individual judgement in deciding whether a patient/family will be approached, which can be influenced by a number of extant factors, including 'cooperativeness,' the perceived likelihood of them agreeing to participate. Recruitment, therefore, is an equity issue - every eligible patient/family should be approached.

Second, it is well established that most of the reasons that marginalized persons are hesitant or even overtly negative toward research is the result of a combination of knowledge concerning well-documented case of abuse at the hands of researchers along with current negative healthcare experiences. These stories are passed through generations and influence both the survivors of Indigenous genocide and subsequent generations. It is incumbent upon researchers to prepare to answer questions about the safety and inclusive practices to support equitable research. We also need to respect patients and families who decline to participate in research.

Different consent paradigms exist in a clinical research context, including waivers of consent, implied consent, and explicit consent. Ethics review boards can take these considerations into account to determine the reasonableness of a proposed consent model.

3.5.3 Comparison with the standard of care. The comparison of ML tools with the current standard is a vital step to ensure that novel interventions actually improve care. However, models are trained from current patterns in data – if biased, it often means that the status quo is likely biased as well. Therefore, the comparison with a given standard must also consider the potential unequal impact on marginalized patients. Often, an unstated presumption with ML bias is that the 'ground truth' (the status quo) is the natural state, which should be replicated faithfully with AI. However, the increasingly well-recognized problems of 'race medicine,' mistreatment of trans and gender diverse persons, neglect of women's health issues, prejudicial assumptions about behaviours among racialized groups, and other systematically biased patterns in medicine complicate this picture.

Vyas and colleagues, for example, offer a series of clinical prediction algorithms in which race is taken into account [100]. One question is whether ML methods may actually improve the current standard by replacing it. Pierson and colleagues [8] demonstrate how clinical knee pain assessment (which is known to be influenced by outdated beliefs about pain tolerance among Black individuals) done by a ML system on the basis of image data results in a higher proportion of Black patients being referred for surgery compared with the status quo. Notably, the model was only assessed retrospectively and thus prospective validity warrants verification..

3.6 Oversight and monitoring

A gap in this framework concerns the ability to prospectively monitor algorithmic performance, an oversight mechanism that is an active area of study for many healthcare institutions [101] [102] [37]. The recommendations themselves may differ depending on the context in which the model is integrated – where some have research teams able to maintain oversight of all clinical algorithms [23] others do not have such resources. It is encouraged that institutions using ML models find a way of prospectively verifying a model's performance and the care delivery to their patients, maintaining an eye toward benefit to patient care, disaggregated to attend to vulnerable patient groups.

Accountability requires that some*one* is responsible to do some*thing*. Defining roles with respect to AI oversight and decision-making is being pursued by many regulatory and professional groups (e.g., Health Education England [103] [104]). Professional colleges can provide such direction to their membership, such as the Royal Australian and New Zealand College of Radiologists (RANZCR) who specify that it is a choice to use AI or not, and one that requires a minimum standard is set such that no patient experiences disadvantage by virtue of its use [105]. Similarly, the Royal College of Physicians and Surgeons of Canada has explored how to support clinicians in navigating an area of regulatory uncertainty while maintaining medicolegal liability for AI use [106].

Some scholars have highlighted that the risks of an uncertain regulatory space are most likely to fall to patients, and particularly those who are marginalized [107]. Through accountability mechanisms such as 'no-fault' compensation models whereby the patient does need to identify the specific agent of harm but can demonstrate that harm arose for the use of an ML tool, some propose that AI oversight can be a matter of collective responsibility [107].

4 DISCUSSION

The JustEFAB guideline represents an initial development of an institutional guideline to address algorithmic bias in healthcare ML. It describes the scope of and process for incorporating ethical principles relevant to bias into the full ML lifecycle and can be used by researchers and developers, ethics review board members, healthcare decision-makers, regulatory bodies, and others who are concerned with the potential for exacerbation of health inequities involving ML. The practices identified in this guideline can form a part of complementary practices to explore model performance.

We strongly advocate for cross-disciplinary capacity building toward a transdisciplinary science of fairness in ML; we hope that JustEFAB can be a step in this direction. We foresee the use of JustEFAB in a number of different ways: 1) encouraging collaboration between ML developers and multidisciplinary scholars in both public and private sectors to develop ML products guided by ethics as a lifecycle approach to guard against unfairness prior to deployment; 2) ethics review boards can use the framework as part of the review process for ML research studies to protect participant wellbeing and justice as part of research ethics oversight; 3) institutional decision-makers can use the framework to vet ML applications which they may be interested in trialing or purchasing locally to protect their interests and those of their patients; 4) clinical users can use the framework to know which questions to ask when collaborating on the development or testing of an ML application.

We also foresee the benefit of translating this framework into a non-expert, accessible format (e.g., a tool or template) to be more broadly available. Clinicians may be faced with questions from patients, families, and communities who wish to understand the safety and vetting process for ML tools used in their care or the care of a loved one.

5 CONCLUSION

Fairness in medicine evolves with social and political circumstances simultaneous to the advancement of scientific knowledge. Assuming that healthcare and science are value-neutral has proven dangerous [5]; the way forward is to embrace values in the choices we make to better the care of others. This guideline provides a starting point for values-based decisions regarding issues of algorithmic bias in healthcare ML. We hope that it proves helpful for ML developers, researchers, clinicians, ethics review boards, healthcare decision-makers, regulatory bodies, and professional organizations. We imagine future work will iterate upon and improve this initial attempt at providing constructive, holistic guidance regarding algorithmic bias for health ML.

ACKNOWLEDGMENTS

We express our gratitude to the multitude of researchers, clinicians, patients, families, community members, and scholars whose thoughts and works have contributed to the thinking embedded in this guideline's development. We are grateful especially to the consultative groups at SickKids who took the time and courage to speak their minds, ask questions, and give us their honest feedback about this difficult issue. The lead author wishes to express gratitude to James Anderson and Caesar Atuire for their careful reading of the paper to provide helpful feedback. Thanks also goes out to Nate Stein at VisualDX for engaging around the VisualDX use case (Appendix B).

REFERENCES

[1] A. Birhane, "Algorithmic injustice: a relational ethics approach," *Patterns,* vol. 2, no. 2, p. 100205, 2021.
[2] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," *Proceedings of Machine Learning Research,* pp. 77-91, 2018.
[3] S. Corbett-Davies and S. Goel, "The measure and mismeasure of fairness: A critical review of fair machine learning," *arXiv preprint arXiv:1808.00023,* 2018.
[4] S. Barocas, M. Hardt and A. Narayanan, "Fairness in machine learning,"," *NIPS tutorial 1,* 2017.
[5] R. Benjamin, "Assessing risk, automating racism," *Science,* vol. 366, no. 6464, pp. 421-422, 2019.
[6] Z. P. B. V. C. &. M. S. Obermeyer, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science,* pp. 447-453, 2019.
[7] I. Y. Chen, P. Szolovits and M. Ghassemi, "Can AI help reduce disparities in general medical and mental health care?," *AMA Journal of Ethics,* vol. 21, no. 2, pp. 167-179, 2019.
[8] E. Pierson, D. M. Cutler, J. Leskovec, S. Mullainathan and Z. Obermeyer, "An algorithmic approach to reducing unexplained pain disparities in underserved populations," *Nature Medicine,* vol. 27, no. 1, pp. 136-140, 2021.
[9] H. Zhang, A. X. Lu, M. Abdalla, M. McDermott and M. Ghassemi, "Hurtful words: quantifying biases in clinical contextual word embeddings," *Proceedings of the ACM Conference on Health, Inference, and Learning,* pp. 110-120, 2020.
[10] L. Seyyed-Kalantari, H. Zhang, M. B. McDermott, I. Y. Chen and M. Ghassemi, "Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations," *Nature Medicine,* vol. 27, no. 12, pp. 2176-2182, 2021.
[11] X. Liu, B. Glocker, M. Melissa, M. Ghassemi, A. K. Denniston and L. Oakden-Rayner, "The medical algorithmic audit," *The Lancet Digital Health,* vol. 2022.
[12] I. D. Raji, A. Smart, R. N. White, M. Mitchell, T. Gebru, B. Hutchinson, J. Smith-Loud, D. Theron and P. Barnes, "Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic audit," *Proceedings of the 2020 conference on fairness, accountability, and transparency,* pp. 33-44, 2020.
[13] L. Oakden-Rayner, W. Gale, T. A. Bonham, M. P. Lungren, G. Carneiro, A. P. Bradley and L. J. Palmer, "Validation and algorithmic audit of a deep learning system for the detection of proximal femoral fractures in patients in the emergency department: a diagnostic accuracy study," *The Lancet Digital Health,* vol. 4, no. 5, pp. e351-e358, 2022.
[14] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. Daumé Iii and K. Crawford, "Datasheets for datasets," *Communications of the ACM,* vol. 64, no. 12, pp. 86-92, 2021.
[15] N. Rostamzadeh, D. Mincu, S. Roy, A. Smart, L. Wilcox, M. Pushkarna, J. Schrouff, R. Amironesei, N. Moorosi and K. Heller, "Healthsheet: development of a transparency artifact for health datasets," *2022 ACM Conference on Fairness, Accountability, and Transparency,* pp. 1943-1961, 2022.
[16] A. Rajkomar, M. Hardt, M. D. Howell, G. Corrado and M. H. Chin, "Ensuring fairness in machine learning to advance health equity," *Annals of internal medicine 169, no. 12 (2018): ,* vol. 169, no. 12, pp. 866-872, 2018.
[17] J. W. Gichoya, L. G. McCoy, L. A. Celi and M. Ghassemi, "Equity in essence: a call for operationalising fairness in machine learning for healthcare," *BMJ health & care informatics,* vol. 28, no. 1, 2021.
[18] S. R. Pfohl, A. Foryciarz and N. H. Shah, "An empirical characterization of fair machine learning for clinical risk prediction," *Journal of biomedical informatics,* vol. 113, p. 103621, 2021.
[19] I. Dankwa-Mullan, E. L. Scheufele, M. E. Matheny, Y. Quintana, W. W. Chapman, G. Jackson and B. R. South, "A proposed framework on integrating health equity and racial justice into the artificial intelligence development lifecycle," *Journal of Health Care for the Poor and Underserved,* vol. 32, no. 2, pp. 300-317, 2021.
[20] I. Y. Chen, E. Pierson, S. Rose, S. Joshi, K. Ferryman and M. Ghassemi, "Ethical machine learning in health care," *Annual review of biomedical data science,* vol. 4, pp. 123-144, 2021.
[21] S. Fazelpour, Z. C. Lipton and D. Danks, "Algorithmic fairness and the situated dynamics of justice.," *Canadian Journal of Philosophy,* vol. 52, no. 1, pp. 44-60, 2022.
[22] M. D. McCradden, S. Joshi, J. A. Anderson, M. Mazwi, A. Goldenberg and R. Zlotnik Shaul, "Patient safety and quality improvement: Ethical principles for a regulatory approach to bias in healthcare machine learning," *Journal of the American Medical Informatics Association,* vol. 27, no. 12, pp. 2024-2027, 2020.
[23] Z. Obermeyer, R. Nissan, M. Stern, S. Eaneff, E. J. Bembeneck and S. Mullainathan, "Algorithmic Bias Playbook," Center for Applied Artificial Intelligence, Chicago Booth, 2021.
[24] R. Schwartz, A. Vassilev, K. Greene, L. Perine, A. Burt and P. Hall, "NIST Special Publication 1270: Towards a Standard for Identifying and Managing Bias in Artificial Intelligen," National Institute of Standards and Technology, US Department of Commerce, 2022.
[25] C. Ewuoso, "An African relational approach to healthcare and big data challenges," *Science and Engineering Ethics,* vol. 27, no. 3, p. 34, 2021.
[26] S. A. Friedler, C. Scheidegger and S. Venkatasubramanian, "On the (im) possibility of fairness," *arXiv preprint arXiv:1609.07236,* 2016.
[27] R. Zemel, Y. Wu, K. Swersky, T. Pitassi and C. Dwork, "Learning fair representations," *Proceedings of Machine Learning Research,* vol. 17, pp. 325-333, 2013
[28] H. Edwards and A. Storkey, "Censoring Representations with an Adversary," *Proceedings of the 4th International Conference on Learning Representations,* 2016.
[29] D. Madras, E. Creager, T. Pitassi and R. Zemel, "Learning Adversarially Fair and Transferable Representations," *PMLR, 10–15 Jul 2018: 3384–93,* vol. 10, pp. 3384-3393, 2018.

[30] B. Kim, H. Kim, K. Kim, S. Kim and J. Kim, "Learning not to learn: Training deep neural networks with biased data," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, p. 9012–20, 2019.
[31] M. Alvi, A. Zisserman and C. Nellåker, "Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings," *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018.
[32] M. D. McCradden, "When is accuracy off-target?," *Translational Psychiatry,* vol. 11, no. 1, p. 369, 2021.
[33] M. D. J. S. M. M. &. A. J. A. McCradden, "Ethical limitations of algorithmic fairness solutions in health care machine learning," *The Lancet Digital Health,* vol. 2, no. 5, pp. e221-e223., 2020.
[34] H. Suresh and J. Guttag, "A framework for understanding sources of harm throughout the machine learning life cycle," *Equity and access in algorithms, mechanisms, and optimization*, pp. 1-9, 2021.
[35] E. J. Topol, "Welcoming new guidelines for AI clinical research," Nature Medicine, vol. 26, no. 9, pp. 1318-1320, 2020.
[36] M. D. McCradden, J. A. Anderson, E. A. Stephenson, E. Drysdale, L. Erdman, A. Goldenberg and R. Zlotnik Shaul, "A research ethics framework for the clinical translation of healthcare machine learning," American Journal of Bioethics, vol. 22, no. 5, pp. 8-22, 2022.
[37] N. I. f. H. a. C. Excellence, "Evidence standards framework for digital health technologies," National Institute for Health and Care Excellence, United Kingdom, 2022.
[38] S. Mohamed, M.-T. Png and W. Isaac, "Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence," Philosophy & Technology , vol. 33, pp. 659-684, 2020.
[39] T. Cole, "The white-savior industrial complex," The Atlantic, 21 March 2012.
[40] J. G. Faulkenberry, A. Luberti and S. Craig, "Electronic health records, mobile health, and the challenge of improving global health," Current problems in pediatric and adolescent health care , vol. 52, no. 2, p. 101111, 2022.
[41] A. Schwab, "Epistemic humility and medical practice: translating epistemic categories into ethical obligations," Journal of Medicine & Philosophy , vol. 37, no. 1, pp. 28-48, 2012.
[42] K. Ndlovu, N. Stein, M. Annechino, M. Molwantwa, M. Monkge, A. Forrestel and V. L. Williams, "Evaluating Feasibility and Acceptance of a Mobile Clinical Decision Support System in Botswana," [Pending].
[43] K. Ndlovu, "Feasibility, acceptance and ethical considerations of a mobile clinical decision support system in Botswana," in Global Forum on Bioethics Research, Cape Town, 2022.
[44] M. Sendak, M. C. Elish, M. Gao, J. Futoma, W. Ratliff, M. Nichols, A. Bedoya, S. Balu and C. O'Brien, "The human body is a black box" supporting clinical decision-making with deep learning," Proceedings of the 2020 conference on fairness, accountability, and transparency, pp. 99-109, 2020.
[45] S. Harding, Feminism and methodology: Social science issues, Indiana University Press, 1987.
[46] A. Wylie, R. Figueroa and S. Harding, "Why standpoint matters," Science and other cultures: Issues in philosophies of science and technology, vol. 1, pp. 26-48, 2003.
[47] S. Harding, Whose science? Whose knowledge?: Thinking from women's lives, Cornell University Press, 1991.
[48] P. H. Collins, "Learning from the outsider within: The sociological significance of Black feminist thought," Social problems, vol. 33, no. 6, pp. s14-s32, 1986.
[49] O. P. Matshabane, L. Mgweba-Bewana, C. A. Atuire, J. de Vries and L. M. Koehly, "Cultural diversity is crucial for African neuroethics," Nature Human Behaviour, vol. 6, no. 9, pp. 1185-1187, 2022.
[50] L. S. Herzog, S. R. Wright, J. J. Pennington and L. Richardson, "The KAIROS Blanket Exercise: Engaging Indigenous ways of knowing to foster critical consciousness in medical education," Medical Teacher, vol. 43, no. 12, pp. 1437-1443, 2021.
[51] R. L. C. P. R. B. C. E. C. Jones, M. Green, T. Huria, K. Jacklin, M. Kamaka, C. Lacey and J. Milroy, "Educating for Indigenous health equity: an international consensus statement," Academic Medicine, vol. 94, no. 4, p. 512–519, 2019.
[52] E. a. M. National Academies of Sciences, "Improving Representation in Clinical Trials and Research: Building Research Equity for Women and Underrepresented Groups," The National Academies Press. , Washington, DC, 2022.
[53] M. D. Kelsey, B. Patrick-Lake, R. Abdulai, U. C. Broedl, A. Brown, E. Cohn, L. H. Curtis, C. Komelasky, M. Mbagwu, G. A. Mensah, R. J. Mentz, A. Nyaku, S. O. Omokaro, J. Sewards and e. al, "Inclusion and diversity in clinical trials: Actionable steps to drive lasting change," Contemporary Clinical Trials, vol. 116, 2022.
[54] D. L. MacLennan, J. L. Plahovinsak, R. J. MacLennan and C. T. Jones, "Clinical Trial Site Perspectives and Practices on Study Participant Diversity and Inclusion," Clinical Pharmacology & Therapeutics, 2022.
[55] J. M. Kahn, D. M. Gray, J. M. Oliveri, C. M. Washington, C. R. DeGraffinreid and E. D. Paskett, "Strategies to improve diversity, equity, and inclusion in clinical trials," Cancer , vol. 128, no. 2, pp. 216-221, 2022.
[56] L. Oakden-Rayner, J. Dunnmon, C. Gustavo and C. Ré, "Hidden Stratification Causes Clinically Meaningful Failures in Machine Learning for Medical Imaging," Proc ACM Conf Health Inference Learn, pp. 151-159, 2020.
[57] I. M. Young, "Social groups in associative democracy," Politics & Society, vol. 20, no. 4, pp. 529-534, 1992.
[58] A. J. London, "Artificial intelligence in medicine: overcoming or recapitulating structural challenges to improving patient care?," Cell Reports Medicine, vol. 3, no. 5, p. 100622, 2022.
[59] S. Ganapathi, J. Palmer, J. E. Alderman, M. Calvert, C. Espinoza, J. Gath, M. Ghassemi and e. al, "Tackling bias in AI health datasets through the STANDING Together initiative," Nature Medicine, vol. 28, pp. 2232-2233, 2022.
[60] R. Daneshjou, K. Vodrahalli, R. A. Novoa, M. Jenkins, W. Liang, V. Rotemberg, J. Ko and e. al., "Disparities in dermatology AI performance on a diverse, curated clinical image set," Science Advances, vol. 8, no. 31, p. eabq6147, 2022.
[61] J. Xu, B. S. Glicksberg, C. Su, P. Walker, J. Bian and F. Wang, "Federated learning for health informatics," Journal of Healthcare Informatics Research, no. 5, pp. 1-19, 2021.
[62] S. Cui, W. Pan, J. Liang, C. Zhang and F. Wang, "ddressing algorithmic disparity and performance inconsistency in federated learning," Advances in Neural Information Processing Systems, no. 34, pp. 26091-26102, 2021.
[63] K. Crenshaw, "Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics," University of Chicago Legal Forum, vol. 1, no. 8, pp. 138-1677 1989.
[64] A. Ghosh, L. Genuit and M. Reagan, "Characterizing Intersectional Group Fairness with Worst-Case Comparisons," Proceedings of Machine Learning Research, vol. 142, pp. 22-34, 2021.
[65] U. Hebert-Johnson, M. Kim, O. Reingold and G. Rothblum, "Multicalibration: Calibration for the (Computationally-Identifiable) Masses," Proceedings of the International Conference on Machine Learning, vol. 80, pp. 1939-1948, 2018.
[66] J. P. Cerdeña, V. Grubbs and A. L. Non, "Racialising genetic risk: assumptions, realities, and recommendations," The Lancet, vol. 400, no. 10368, pp. 2147-2154, 2022.
[67] J. M. Cénat, "Who is Black? The urgency of accurately defining the Black population when conducting health research in Canada," CMAJ , vol. 194, no. 27, pp. E948-E949, 2022.
[68] C. J. P. Harrell, T. I. Burford, B. N. Cage, T. McNair Nelson, S. Shearon, A. Thompson and S. Green, "Multiple pathways linking racism to health outcomes," Du Bois review: Social Science Research on Race, vol. 8, no. 1, 2011.
[69] L. Pereira, L. Mutesa, P. Tindana and M. Ramsay, "African genetic diversity and adaptation inform a precision medicine agenda," Nature Reviews Genetics , vol. 22, no. 5, pp. 284-306, 2021.
[70] S. S. Richardson, "Sex Contextualism," Philosophy, Theory, and Practice in Biology, vol. 14, no. 2, 2022.
[71] M. DiMarco, H. Zhao, M. Boulicault and S. S. Richardson, "Why "sex as a biological variable" conflicts with precision medicine initiatives," Cell Reports Medicine, vol. 3, no. 4, p. 100550, 2022.
[72] S. C. Chang and A. A. Singh, A clinician's guide to gender-affirming care: Working with transgender and gender nonconforming clients, New Harbinger Publications, 2018.
[73] A. Suess Schwend, "Trans health care from a depathologization and human rights perspective," Public Health Reviews, vol. 41, no. 1, pp. 1-17, 2020.
[74] L. Kcomt, "Profound health-care discrimination experienced by transgender people: rapid systematic review," Social Work in Health Care, vol. 58, no. 2, pp. 201-219, 2019.
[75] C. A. Kronk, A. R. Everhart, F. Ashley, H. M. Thompson, T. E. Schall, T. G. Goetz and e. al, "Transgender data collection in the electronic health record: current concepts and issues," Journal of the American Medical Informatics Association, vol. 29, no. 2, pp. 271-284, 2022.
[76] K. Albert and M. Delano, "Sex trouble: Sex/gender slippage, sex confusion, and sex obsession in machine learning using electronic health records," Patterns, vol. 3, no. 8, p. 100534, 2022.
[77] M. Bernhardt, C. Jones and B. Glocker, "Potential sources of dataset bias complicate investigation of underdiagnosis by machine learning algorithms," Nature Medicine , vol. 28, no. 6, pp. 1157-1158, 2022.
[78] P. Mukherjee, T. C. Shen, J. Liu, T. Mathai, O. Shafaat and R. M. Summers, "Confounding factors need to be accounted for in assessing bias by machine learning algorithms," Nature Medicine, vol. 28, no. 6, pp. 1159-1160, 2022.
[79] A. F. Cooper and E. Abrams, "Emergent unfairness in algorithmic fairness-accuracy trade-off research," Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, pp. 46-54, 2021.
[80] P. J. Embi, "Algorithmovigilance—advancing methods to analyze and monitor artificial intelligence–driven health care for effectiveness and equity," JAMA Network Open, vol. 4, no. 4, pp. e214622-e214622, 2021.
[81] J. W. Gichoya, I. Banerjee, A. R. Bhimireddy, J. L. Burns, L. A. Celi, L.-C. Chen, R. Correa, N. Dullerud, M. Ghassemi, S.-C. Huang and e. al, "AI recognition of patient race in medical imaging: a modelling study," The Lancet Digital Health, vol. 4, no. 6, pp. e406-e414, 2022.
[82] J. J. van Delden and G. J. van Thiel, "Reflective equilibrium as a normative-empirical model in bioethics," Reflective equilibrium: Essays in honour of Robert Heeger, pp. 251-259, 1998.

[83] J. Rawls, A Theory of Justice, 2nd Edition, Cambridge, MA: Harvard University Press, 1971.
[84] A. Birhane, P. Kalluri, D. Card, W. Agnew, R. Dotan and M. Bao, "The values encoded in machine learning research," arXiv preprint, 2021.
[85] I. J. Chasnoff, H. J. Landress and M. E. Barrett, "The prevalence of illicit-drug or alcohol use during pregnancy and discrepancies in mandatory reporting in Pinellas County, Florida."," New England Journal of Medicine, vol. 322, no. 17, pp. 1202-1206, 1990.
[86] N. C. Dlova, R. Gathers, J. Tsoka-Gwegweni and R. Hift, "Skin cancer awareness and sunscreen use among outpatients of a South African hospital: need for vigorous public education," South African Family Practice , vol. 60, no. 4, p. 132, 2018.
[87] M. D. McCradden, J. A. Anderson and M. D. Cusimano, "When is death in a child's best interest?: examining decisions following severe brain injury," JAMA Pediatrics , vol. 173, no. 3, pp. 213-214, 2019.
[88] Y. Park, J. Hu, M. Singh, I. Sylla, I. Dankwa-Mullan, E. Koski and A. K. Das, "Comparison of methods to reduce bias from clinical prediction models of postpartum depression," JAMA Network Open, vol. 4, no. 4, pp. e213909-e2139, 2021.
[89] J. Wiens, S. Saria, M. Sendak, M. Ghassemi, V. X. Liu, F. Doshi-Velex, K. Jung and e. al, "Do no harm: a roadmap for responsible machine learning for health care," Nature Medicine, vol. 25, no. 9, pp. 1337-1340, 2019.
[90] M. D. McCradden, E. A. Stephenson and J. A. Anderson, "Clinical research underlies ethical integration of healthcare artificial intelligence," Nature Medicine , vol. 26, no. 9, pp. 1325-1326, 2020.
[91] M. D. McCradden, "A silent trial is critical to accountable and justice-promoting implementation of artificial intelligence in healthcare," in Global Forum of Bioethics Research, Cape Town, 2022.
[92] A. J. London, For the Common Good: Philosophical Foundations of Research Ethics, Oxford University Press, 2021.
[93] V. Mahajan, V. Kumar Venugopal, M. Murugavel and H. Mahajan, "The algorithmic audit: working with vendors to validate radiology-AI algorithms—how we do it," Academic Radiology, vol. 27, no. 1, pp. 132-135, 2020.
[94] X. Liu, S. Cruz Rivera, D. Moher, M. J. Calvert, A. K. Denniston, H. Ashrafian, A. L. Beam and e. al., "Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension," Nature Medicine, vol. 26, pp. 1364-1374, 2020.
[95] S. Cruz Rivera, X. Liu, A.-W. Chan, A. K. Denniston, M. J. Calvert, H. Ashrafian, A. L. Beam and e. al, "Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension," Nature Medicine, vol. 26, pp. 1351-1363, 2020.
[96] S. Tonekaboni, G. Morgenshtern, A. Assadi, A. Pokhrel, X. Huang, A. Jayarajan, R. Greer and e. al, "How to validate Machine Learning Models Prior to Deployment: Silent trial protocol for evaluation of real-time models at ICU," Conference on Health, Inference, and Learning, pp. 169-182, 2022.
[97] B. Vasey, M. Nagendran, B. Campbell, D. A. Clifton, G. S. Collins, S. Denaxas, A. K. Denniston and e. al., "Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI," Nature Medicine, vol. 28, no. 5, pp. 924-933, 2022.
[98] M. Madden and E. Ewen Speed, "Beware zombies and unicorns: toward critical patient and public involvement in health research in a neoliberal context," Frontiers in Sociology, p. 7, 2017.
[99] E. Melnick, L. Dyrbye, C. Sinsky, M. Trockel, C. West, L. Nedelec and e. al, "The association between perceived electronic health record usability and professional burnout among US physicians," Mayo Clinic Proceedings, vol. 95, no. 3, pp. 476-487, 2020.
[100] E. Lett, E. Asabor, S. Beltran and A. M. A. O. A. Cannon, "Conceptualizing, contextualizing, and operationalizing race in quantitative health sciences research," Annals of Family Medicine, vol. 20, pp. 157-163, 2022.
[101] N. C. f. t. P. o. H. S. o. B. a. B. Research, "The Belmont report: Ethical principles and guidelines for the protection of human subjects of research," U.S. Department of Health and Human Services, 1979.
[102] K. Spector-Bagdady, S. Tang, S. Jabbour, W. N. Price, A. Bracic, M. S. Creary, S. Kheterpal, C. M. Brummett and J. Wiens, "Respecting autonomy and enabling diversity: the effect of eligibility and enrollment on research data demographics," Health Affairs, vol. 40, no. 12, 2021.
[103] D. A. Vyas, L. G. Eisenstein and D. S. Jones, "Hidden in plain sight—reconsidering the use of race correction in clinical algorithms," New England Journal of Medicine , vol. 383, no. 9, pp. 874-882, 2020.
[104] M. Sendak, G. Sirdeshmukh, T. Ochoa, H. Premo, L. Tang, K. Niederhoffer, S. Reed and e. al, "Development and Validation of ML-DQA–a Machine Learning Data Quality Assurance Framework for Healthcare," arXiv preprint, 2022.
[105] S. G. Finlayson, A. Subbaswamy, K. Singh, J. Bowers, A. Kupke, J. Zittrain, I. S. Kohane and S. Saria, "The clinician and dataset shift in artificial intelligence," New England Journal of Medicine 385, no. 3 (2021): 283, vol. 385, no. 3, pp. 283-286, 2021.
[106] M. Nix, G. Onisforou and A. Painter, "Understanding healthcare workers' confidence in AI: Report 1 of 2," NHS AI Lab & Health Education, 2022.
[107] M. Nix, G. Onisforou and A. Painter, "Developing healthcare workers' confidence in AI: Report 2 of 2," NHS AI Lab & Health Education England, 2022.
[108] A. I. A. Committee, "Standards of Practice for Artificial Intelligence," Royal Australian and New Zealand College of Radiologists (RANZCR), Sydney, NSW, 2020.
[109] R. K. Reznick, K. Harris, T. Horsley and M. Skeikh Hassani, "Task Force Report on Artificial Intelligence and Emerging Digital Technologies," Royal College of Physicians and Surgeons of Canada, 2020.
[110] J. Yoshikawa, "Sharing the costs of artificial intelligence: Universal no-fault social insurance for personal injuries," Vand. J. Ent. & Tech. L., vol. 21, p. 1155, 2018.
[111] P. Schulam and S. Saria, "Can you trust this prediction? Auditing pointwise reliability after learning," Proc Mach Learn Res, vol. 89, pp. 1022-1031, 2019.
[112] T. Cole, "The white-savior industrial complex," The Atlantic, 21 March 2012.
[113] A. Schwab, "Epistemic humility and medical practice: translating epistemic categories into ethical obligations," Journal of Medicine & Philosophy , vol. 37, no. 1, pp. 28-48, 2012

APPENDIX A: JUSTEFAB FRAMEWORK

Table 1: JustEFAB Framework for addressing fairness in health ML tools

STAGE 1A: DESIGN AND DEVELOPMENT (3.1)		
Preparation and Conceptualization (3.1.1)	Literature review and consultation	A priori group identification Problem formulation Consultation
Dataset Inclusivity (3.1.2)	Knowing the inputs and understanding their context	Representation and labels Dataset reflexivity Analytic plan
Algorithmic validation (3.1.3)	Exploring the fairness distribution of the candidate model	Test and compare performance among subgroups Algorithmic fairness methods Testing for post hoc grouping (hidden stratification)
STAGE 1B: ETHICAL DECISION-MAKING (3.2)		
Ethical Decision-Making (3.2)	Selecting a fairness strategy to guide implementation	Identifying biases Reflective equilibrium Selection of prioritized value
STAGE 2: SILENT TRIAL AND CLINICAL VALIDATION OF FAIRNESS (3.3)		
Clinical Performance (3.3.1)	Establishing the on-the-ground performance of the algorithm overall and with respect to relevant patient subgroups	Clinical accuracy Characterize the performance across relevant subgroups Auditing Revision
Human Factors (3.3.2)	Applying psychological principles to the engineering and design of ML tools	Human-centred design Respect for persons Ethical use considerations
STAGE 3: PROSPECTIVE CLINICAL EVALUATION (3.4)		
Diversity (3.4.1)	Inclusive research design to enable diverse participation	
Consent Considerations (3.4.2)	Thinking carefully about the ethical justifiability of consent paradigms	
Comparison to Status Quo (3.4.3)	Considering the current evidence with respect to a health disparity and how to move toward an ideal state	
STAGE 4: OVERSIGHT AND MONITORING (3.5)		
Patient safety and Accountability	Taking ownership of ML system performance, oversight, and decision-making; identifying parties who take responsibility for algorithmic fairness decisions; identifying patient safety mechanisms	

APPENDIX B: VISUALDX AS AN EXTERNAL USE CASE

The number of dermatology specialists in Botswana's public health sector has varied from none to most recently 2 full time MOH employees and three contract specialists from Cuba. However, the demand for dermatology care continues to be much higher than can be provided by the current specialists resulting in six or more months of waiting times for appointments. This shortage of dermatology specialists in Botswana necessitated efficient use of the limited resources and continuous empowerment of those commonly engaged in the management of prevalent skin conditions. It further suggests a critical need for a clinical decision support system (CDSS) to ameliorate current challenges.

In 2020, the University of Botswana (UB) collaborated with VisualDx on a research study funded by the Bill & Melinda Gates Foundation (grant number INV003773) to assess the feasibility of VisualDx usage in patient care settings in Botswana and also gather feedback to inform further improvements of the platform. Prior to VisualDx implementation in Botswana, research ethical clearance was sought through UB and the Ministry of Health (MOH). A total of 20 dermatology clinics in Botswana participated and these were nominated by the Gaborone District Health Management Team (DHMT). The DHMT is a local authority under MOHW tasked with overlooking management and staffing of primary care clinics. Two VisualDx employees supported the research project by attending weekly update meetings and also supporting virtual user training. No feature modifications were introduced on the VisualDx platform prior to implementation in Botswana and product intellectual property rights remained with VisualDx.

VisualDx has over 20 years of experience in supporting healthcare providers with their clinical decision making. It employs over 70 full-time team members all dedicated to maintaining accurate, up to date content with user friendly functionality. The platform has become a standard professional resource at more than 2,300+ universities, hospitals, and clinical sites globally. It combines expert knowledge, problem-oriented search, the world's best curated medical image library, and technology to support differential diagnosis,

treatment recommendations, and patient education. VisualDx is available on the web, native iOS and Android applications and most recently includes off-line capability on Android devices. VisualDx has the potential to contribute to increased provider confidence and a reduction in diagnostic errors in primary care settings. The platform combines machine learning algorithms and vision science with a structured clinical knowledge base to allow non-specialist healthcare providers to capture patient-specific findings, build custom differentials, and view images and treatment recommendations. The DermExpertTM feature in VisualDx uses a Convolutional Neural Network (CNN) to estimate diagnosis and lesion categories from an input image. CNNs are data-driven models that require a large dataset of labeled pairs to train and validate.

Application of JustEFAB

The model's task is one that is of value to Botswanans, participation relied on informed consent, and its use was approved by local institutional review boards [108]. Using JustEFAB could have prompted the a priori identification of the skin pigmentation ranges where model performance was lower (3.1.1, 3.1.3), facilitated by a local silent trial (3.3.1) prior to the evaluation of VisualDX in the target population. By including new case examples from the Botswanan population, the model's fairness parameters were improved [108].

This example can be considered one of prioritizing formal equality by increasing representation in the training dataset (3.2.3), wherein the choice to improve the model's training by including new cases improved its performance overall. The improvement facilitates the equal treatment of individuals to improve detection of skin cancers in the Botswanan population [108]. In this case, all images are treated similarly in terms of the computational processing. One can imagine that once the model achieves comparable performance across all levels of skin pigmentation, the guiding value would shift toward predictive accuracy.

While an important step to improve care, Ndlovu notes that limitations in infrastructure and community acceptance may limit VisualDX's benefits [108] [109]. Additionally, there may be limitations with respect to access to technologies as well as appreciation of skin cancer risk that would prompt an individual to seek access to the tool. Complementary efforts (3.3.2) could entail public health awareness and messaging to individuals to improve detection, motivated by VisualDX's placement as a facilitator of care access. These efforts highlight how technical notions of fairness are maximized when coupled with larger efforts toward fair access and treatment in healthcare.

APPENDIX C: EXTENDED METHODS

2.1 Framework development, refinement, and validation

2.1.1 Conceptualization and development of JustEFAB

The initial concept was identified by ethics consultations at the lead institution brought to some of these authors at different stages (e.g., research ethics consultations, general advice for design of ML products, and during the validation of a specific model). The need for an institutional approach was apparent by the increasing number of consults, which aligned with the general recognition in the ML community that fairness issues in healthcare ML pose a serious concern for beneficial integration. By drawing from these initial consultations, we identified a core set of ethical principles, moral theories, and methodologies relevant to fairness issues arising at SickKids. These included drawing from local legal standards, institutional policies, professional practices guidelines, as well as bioethics and paediatric bioethics literature broadly. The application of these sources to the specific use cases was the substrate for developing the guideline.

For these consultations, we stayed abreast of relevant developments in the fair ML field keeping a constant eye out toward organizing frameworks that provided guidance. We identified a core set that informed the development of the guideline based on their relevance to medical ethics [24] [23] [34] [20] [19]. Most influential among these was the Algorithmic Bias Playbook [23], which provided a starting point for the process laid out in this framework. In applying this Playbook to our own use cases, we identified areas which required additional practices or knowledge to meet the requirements of the ethical principles we laid out as relevant to integration in a healthcare institution.

Finally, this initial framework was developed predominantly by individuals knowledgeable in ML, meaning that its relevance to those outside of this circle would be limited. Once the draft guideline was finalized, we took it to the relevant groups at SickKids who typically are consulted for policy advancement, and undertook more in-depth consultations with each group housed within the Equity, Diversity, and Inclusion (EDI) Network at our institution (described below). Additional consultations were undertaken at the request of any individual who wanted to engage with the guideline development, which included clinicians, scientists, and administrative professionals.

2.1.2 Consultative groups and consultation process

After identifying the constellation of practices to characterize and address fairness in ML applications, we considered to what end these methods would be applied. Through an inductive process, the ethical decision-making framework was developed as an initial proposal before consultation was sought. The lead author (MDM) conducted consultations with equity-deserving groups at SickKids in partnership with the EDI Executive Lead and co-author (TG). Groups included: the EDI Steering Committee, the Bioethics Department, SickKids Black Caucus, Indigenous Health Council, the Children's Council, the Family-Centred Care Advisory Council, 2SLGBTQIA+ Steering Committee, SickKids Gender Clinic Steering Committee, and the Accessibility and Inclusion Committee. Consultation involved a brief presentation of the rationale for the framework and the overview of the guideline before taking questions and specific feedback around perceptions of competing fairness definitions, justifications for model design choices, preferences for adjunct supports for ML interpretation and use, and other feedback on the concepts and content of the guideline. Each group was provided with the guideline in advance of each meeting. Feedback was supported through both identifiable and anonymous means to improve individuals' comfort with providing honest and candid feedback. All individuals on each committee/group had the opportunity to review the drafted framework in full and provide either direct feedback to the lead author or aggregated, anonymized feedback through their committee/group Chair. We again stress that our view of the consultative process is not a 'one-and-done' endeavour, and maintain a consistent communication with these

groups as the guideline continues to be refined and utilized. We also stress that it does not replace the need for consultation around specific ML tools.

*2.1.3 Incorporation of feedback from consultative groups*Feedback from groups was incorporated across the entire framework. The majority of comments related to the need to go beyond a fair ML system and signal the need for users (typically, clinicians) to engage with anti-racist, anti-oppression, and gender inclusive care practices, in addition to the ongoing prioritization of core healthcare values such as patient autonomy, accessibility, and justice. Without this, consultants expressed a great deal of skepticism that this framework would bring about significant improvements to care, even if the algorithms themselves were developed and integrated in an ethical way. Additionally, consultants signaled the importance of data inclusivity – an issue which is not directly addressed through this framework, though is something advocated in (3.1.2) – as being essential to feeling included in the scientific enterprise and being 'seen.' At the same time, many expressed skepticism and sometimes rejection of data about themselves being collected. The common concern among these individuals was not knowing "what SickKids is going to end up doing with that information" (i.e., selling data, sharing with external parties without the data owner's knowledge, etc). A further few reported to us that the resources going to AI are disproportionate to other areas of care have far more urgent concerns. This is an important point, in our view, and resulted in more effort toward establishing the social and scientific value of any ML model development effort [110]. A final remark relevant to this guideline is that consultants highlighted how historical efforts to be more inclusive have paradoxically resulted in harms. For example 'White Saviorism' [111] is well documented in global health research, including in informatics and mobile health technologies [112]. Our guiding framework of standpoint theory stresses the importance of avoiding a deficit-based lens when regarding minoritized groups and instead adopting an epistemically humble approach to collaboration in recognition of the knowledge held by a multitude of stakeholders [113].

Recognizing that the problem of algorithmic bias is one that is apparent across the globe, as a next step we reached out to collaborators beyond our institution to assess the suitability, adaptability, fit, and relevance of the guideline. Collaborators provided additional content and insight to the guideline to improve its generalizability. We assessed the potential applicability of JustEFAB to the development and testing of VisualDX in Botswana (KN) to demonstrate how application of the framework into the design and development of an algorithm can improve the fair and ethical integration [109] [108].